Understanding of Navy Technical Language via Statistical Parsing

Neil C. Rowe

Code CS/Rp, 833 Dyer Road
Department of Computer Science
Naval Postgraduate School
Monterey, CA USA 93943
ncrowe@nps.navy.mil

Abstract

A key problem in indexing technical information is the interpretation of technical words and word senses, expressions not used in everyday language. This is important for captions on technical images, whose often pithy descriptions can be valuable to decipher. We describe the natural-language processing for MARIE-2, a natural-language information retrieval system for multimedia captions. Our approach is to provide general tools for lexicon enhancement with the specialized words and word senses, and to learn word usage information (both on word senses and word-sense pairs) from a training corpus with a statistical parser. Innovations of our approach are in statistical inheritance of binary co-occurrence probabilities and in weighting of sentence subsequences. MARIE-2 was trained and tested on 616 captions (with 1009 distinct sentences) from the photograph library of a Navy laboratory. The captions had extensive nominal compounds, code phrases, abbreviations, and acronyms, but few verbs, abstract nouns, conjunctions, and pronouns. Experimental results fit a processing time in seconds of $0.0858n^{2.876}$ and a number of tries before finding the best interpretation of $1.809n^{1.668}$ where n is the number of words in the sentence. Use of statistics from previous parses definitely helped in reparsing the same sentences, helped accuracy in parsing of new sentences, and did not hurt time to parse new sentences. Word-sense statistics helped dramatically; statistics on word-sense pairs generally helped but not always.

1. Introduction

Our MARIE project has been investigating information retrieval of multimedia data by emphasizing caption processing. Although media content analysis such as image processing reduces the need to examine captions, caption processing can be much faster since captions summarize important content in an often-small number of words. Checking captions before retrieving media data can rule out bad matches quickly, and captions can provide information not in the media like the date or names of people in a photograph.

Some natural-language processing of captions is necessary for high query recall and precision. Processing must determine the word senses and how the words relate to get beyond the well-known limits of keyword matching (Krovetz and Croft, 1992). This is a challenge for specialized technical dialects. Automatic indexing for them could have a high payoff if few people currently can understand the captions. But linguistically such

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2004		**00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Understanding of Navy Technical Language via Statistical Parsing	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Department of Computer Science,833 Dyer Road,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**36**	

dialects offer (1) unusual words, (2) familiar words in unusual senses, (3) code words, (4) acronyms, and (5) new syntactic features. It is not cost-effective to hand-code all such specifics for every technical dialect. We need to infer most of them from analysis of a representative technical corpus. And then to handle the unusual usage in the dialects, we should mix traditional symbolic parsing with probabilistic ranking from statistics.

While the MARIE project is intended for multimedia information retrieval in general, we have used as testbed the Photo Lab of the Naval Air Warfare Center (NAWC-WD), China Lake, California USA. This is a library of about 100,000 pictures with about 37,000 captions. The pictures cover all activities of the center, including pictures of equipment, tests of equipment, administrative documentation, site visits, and public relations. With so many pictures, many incomprehensible to ordinary people, captions are indispensable to find anything. But the existing computerized keyword system for finding pictures from their captions is unhelpful, and is mostly ignored by personnel.

(Guglielmo and Rowe, 1996) reports on MARIE-1, a prototype system that we developed for NAWC-WD, a system that appears more in the direction of what users want. MARIE-1 followed traditional methods of natural-language processing for information retrieval (Grosz et al, 1987; Rau, 1988; Sembok and van Rijsbergen, 1990) using hand-coded lexicon information. But MARIE-1 took a man-year to construct and only handled 217 pictures (averaging 20 words per caption, and captions were occasionally corrected) from the database, and its handling of unrestricted queries about the data is poor. To do better, MARIE-2 uses statistical parsing and a number of training methods. MARIE-2 also took a man-year of effort, but handles 616 mostly-unedited captions at a higher rate of accuracy. Development encountered some interesting problems, and provides a good test of the application of statistical parsing ideas to an unfiltered real-world dialect. MARIE-2's parser is implemented in semi-compiled Quintus Prolog and took 5600 lines of source code to specify.

2. Example captions

To illustrate the problems, here are example captions from NAWC-WD. All are single-case.

an/apq-89 xan-1 radar set in nose of t-2 buckeye modified aircraft bu# 7074, for flight evaluation test. 3/4 overall view of aircraft on runway.

This is typical of many captions: two noun phrases terminated with periods, where the first describes the photographic subject and the second describes the picture itself. Also typical are the complex nominal compounds, "an/apq-89 xan-1 radar set" and "t-2 buckeye modified aircraft bu# 7074". Domain knowledge is necessary to recognize "an/apq-89" as a radar type, "xan-1" a version number for that radar, "t-2" an aircraft type, "buckeye" a slang additional name for a T-2, "modified" a conventional adjective, and "bu# 7074" as an aircraft code ID. Note that several words here change meaning when they modify other words. Thus the nonsyntactic approach of the indexing system of (Silvester et al, 1994) for a similar domain has limitations.

program walleye, an/awg-16 fire control pod on a-4c bu# 147781 aircraft, china lake on tail, fit test. 3/4 front overall view and closeup 1/4 front view of pod.

This illustrates some common patterns. "A-4c bu# 147781" is in a common form of <equipment-type> <prefix-code> <code-number>, "an/awg-16 fire control pod" is in a form of <equipment-name> <equipment-purpose> <equipment-type>, and "3/4 front overall view" is in a form of <view-qualifier> <view-qualifier> <view-type>.

graphics presentation tid progress 76. sea site update, wasp head director and hawk screech/sun visor radars. top portion only, excellent.

This illustrates the need for domain-dependent statistics on word senses. Here "wasp", "hawk", "screech", and "sun visor" should not be interpreted in their common English word senses, but as equipment terms. Furthermore, "progress 76" means "progress in 1976", "excellent" refers to the quality of the picture, the "head director" is not a person but a guidance system, and the "sea site" is a dry lakebed flooded with water to a few inches.

aerial low oblique, looking s from inyodern rd at main gate down china lake bl to bowman rd. on l, b to t, water reservoirs, trf crcl, pw cmpnd, vieweg school, capehart b housing, burroughs hs, cimarron gardens, east r/c old duplex stor. lot. on r, b to t, trngl, bar s motel, arrowsmith, comarco, hosp and on to bowman rd.

This illustrates abbreviations and misspellings in the captions. "Trf crcl" is "traffic circle", "trngl" is "triangle", "capehart b" is "capehart base", but "b to t" is "bottom to top". "Vieweg" which looks like a misspelling is actually a person name, but "inyodern" should be "inyokern", a nearby town.

In general, most semantic associations and even many syntactic rules in this dialect exhibit quite different frequencies of use compared to everyday English. Table 1 shows some example syntax rules and their observed frequencies in our training and test captions. Also, words are generally less ambiguous than in everyday English: Of the 1858 word senses used in the first three caption sets, 1679 were the only sense used of their word. Nonetheless, the word senses used are often not the most common in standard English, and many ambiguities of word relationships must be resolved.

3. The lexicon

Creating the full synonym list, type hierarchy, and part hierarchy for applications of the size of the NAWC-WD database (29,082 distinct words in 36,191 captions) is considerable work. Fortunately, common words are covered already by Wordnet (Miller et al, 1990), a large thesaurus system that includes synonym, type, and part information, plus rough word frequencies and morphological processing. Wordnet provided basic information for 6,729 words in the NAWC-WD captions (with about 24,000 word senses).

Table 1: Example syntactic rules with their frequency in the output of the parser for all captions studied.

Rule	Frequency	Example
adj2 + ng = ng	2551	"Navy" + "aircraft"
b_prtp + np = prtp2	122	"testing" + "the seat"
art2 + ng = np	288	"the" + "naval aircraft"
adv + participle = a_prtp	28	"just" + "loaded"
noun + numeric = ng	81	"test" + "0345"
timeprepx + np = pp	82	"during" + "the test"
locprepx + np = pp	710	"on" + "the ground"
miscprepx + np = pp	654	"with" + "instrument pod"
np + pp = np	1241	"Navy aircraft" + "during testing"
np + prtp = np	306	"a crewman" + "loading the pod"
vg + np = vp2	53	"loads" + "the instrument pod"
np + vp = snt	53	"a crewman" + "loads the pod"
vp2 + pp = vp2	25	"loads" + "on the aircraft"
adv + pp = pp	24	"just" + "below the aircraft"
conj + np = cj_np	167	"and" + "aircraft"
np + cj_np = np	195	"sled" + "and dummy"
np + c_aps = np	60	"the aircraft" + ", F-18"
ng + aps = ng	195	"aircraft" + "(F-18)"
np + aps = np	54	"the aircraft" + "(F-18)"
np + c_np = np	55	"the aircraft" + ", the F-18"
np_c + np = np	57	"sled," + "dummy" [comma fault]
prtp2 + pp = prtp2	155	"just loaded" + "on aircraft"
infinmarker + vp = ip	19	"to" + "load"
conj + vp = cj_vp	3	"and" + "loads the aircraft"
snt + cj_snt = snt	3	"crewmen load" + "and officer directs"

The remaining words which were handled in several different ways (see Table 2). There are many words with defined code prefixes like "f-" in "f-18" for fighter aircraft, or like "es" in "es4522" for test numbers. We have several hundred rules for special formats including multiword ones, which interpret "bu# 462945" as an aircraft identification number by its front word, "sawtooth mountains" as mountains by its tail word, "02/21/93" as a date, "10-20m" as a range of meters, "visit/dedication" as a conjunction, and "ship-loading" as a noun-gerund equivalent of an adjective. Misspellings and abbreviations were obtained mostly automatically, with human post-checking, using the methods described in (Rowe and Laitinen, 1995). Lexical ellipsis (e.g. "356" after "LBL 355") is also covered. Other important classes of words are technical words from MARIE-1,

morphological variants on known words, numbers, person names, place names, and manufacturer names. 1700 words needed explicit definition by us in the form of part of speech and superconcept or synonym. The remaining unclassified words are assumed to be equipment names, a usually safe assumption. The effort for lexicon-building was relatively modest (0.3 of a man-year) thanks to Wordnet, which suggests good portability. Wordnet also provided us with 15,000 synonyms for the words in our lexicon, and we provided additional synonyms for technical word senses. For each set of synonyms, we picked a "standard synonym". Pointers go from all other synonyms to the standard synonym, the only synonym for which detailed lexical information like superconcepts is kept.

We put all lexicon information in Prolog format. For instance:

d(aircraft, noun, 1, [vehicle-1], [plane-6, autogiro, autogyro, gyroplane, 'lighter-than-air craft'-1, drone-3, glider-1, chopper-2], ['aircraft engine', bay-6, cockpit-2, cabin-2, 'fuel gauge', 'fuel indicator', frame-4], [fleet-1]).

This says that "aircraft" in noun sense 1 is a kind of vehicle in sense 1; the last three lists represent the primary subcategories of aircraft, parts of aircraft, and wholes containing an aircraft. Words with sense numbers are cross-references to other words in the lexicon.

Then the meaning assigned to a noun or verb in a caption is that of an instance of its associated type, and other parts of speech correspond to properties or relationships of a types. Most words of the input caption map to a single two-argument predicate expression in the semantic representation ("meaning list") of the caption. For instance, "Navy aircraft on runway" has the meaning list:

[a_kind_of(v3,aircraft-1), owner(v3,'USN'-1), over(v3,v5), a_kind_of(v5,runway-1)]

where v3 and v5 are unnamed instances, and the numbers after the hyphen are word sense numbers.

4. Parsing

We chose to use a simple grammar and relatively simple semantic rules, to see how far we could rely on statistics in lieu of subtler distinctions (an idea similar to that of (Basili et al, 1992)). For instance, the NAWC-WD corpus frequently has a type of aircraft followed by a "bureau number" code, but we handle this with only a general rule for nominal compounds of physical objects followed by names. Our grammar has 217 syntax rules, 185 binary (two-term) and 32 unary (one-term), and 71 of the binary rules are context-sensitive. The context-sensitivity is modest and unnecessary for correct parsing, but helps efficiency on long sentences. For instance, an appositive that starts with a comma must be followed by comma except at the end of a sentence, and a modifying participial phrase followed by a comma and a noun phrase can only occur at the front of a sentence.

Table 2: Statistics on the MARIE-2 lexicon for the NAWC-WD captions after handling the first caption set; subsequent sets added only 298 new word senses.

Description	Count
Number of captions	36,191
Number of words in the captions	610,182
Number of distinct words in the captions	29.082
Subset having explicit entries in Wordnet	6,729
Number of these for which a preferred alias is given	1,847
Number of word senses given for the Wordnet words	14,676
Subset with definitions reusable from MARIE-1	770
Subset with definitions written explicitly for MARIE-2	1,763
Subset that are morphological variants of other known words	2,335
Subset that are numbers	3,412
Subset that are person names	2,791
Subset that are place names	387
Subset that are manufacturer names	264
Subset that have unambiguous defined-code prefixes	3,256
Unambiguous defined-code prefixes in these	947
Subset that are other identifiable special formats	10,179
Subset that are identifiable misspellings	1,174
Misspellings found automatically of these	713
Subset that are identifiable abbreviations	1,093
Abbreviations found automatically of these	898
Remaining words, assumed to be equipment names	1,876
Explicitly used Wordnet alias facts of above Wordnet words	20,299
Extra alias senses added to lexicon beyond caption vocabulary	9,324
Explicitly created alias facts of above non-Wordnet words	489
Other Wordnet alias facts used in simplifying the lexicon	35,976
Extra word senses added to lexicon beyond caption vocabulary	7,899
Total number of word senses handled (including related superconcepts, wholes, and phrases)	69,447

Binary rules have associated semantic rules (139 in all including one default rule) that check semantic consistency and assemble the combined meaning list. 14 of the semantic rules were specific to the dialect. Additional rules address three constructs especially common in technical description: nominal compounds, appositives, and prepositional phrases. The rules for nominal (noun-noun) compounds cover 62 combinations like type-subtype ("f-18 aircraft"), type-part ("f-18 wing"), owner-object ("navy f-18"), object-action ("f-18 takeoff"), action-object ("training f-18"), action-location ("training area"), object-concept ("f-18 project"), and type-name ("f-18 harrier"). Rules for appositives cover 27 analogous cases and some others like object-action ("wings (folded)"). Rules for prepositional phrases check case compatibility of the preposition with both subject and object. For instance, the object of the location-preposition meaning of "in" could only be a location, event, range, or view; its subject could be only a location or event.

We use a kind of bottom-up chart parser (Charniak, 1993, Chapter 6). We work separately on each sentence of a caption. The most likely interpretation for each word, using word-sense statistics (see next section), is entered in the chart. No initial part-of-speech tagging (Brill, 1995) or initial sense disambiguation (Leacock, Chodorow, and Miller, 1998) from context is done, as this is done indirectly later in successful phrase constructions. We then do a branch-and-bound search in which the highest-rated unexamined word sense or phrase interpretation in the chart is selected at each step. All word and phrase interpretations that adjoin it (without gaps) to the left and right are considered for combination, and checked against grammar rules; if successful, they are checked against semantic rules; and then if successful are rated and added to the chart. A sentence interpretation is a chart entry that covers all the words of the sentence and has the grammatical category "caption". Sentence interpretations are presented to a human trainer for approval. Upon acceptance, conjunct simplification and anaphoric-reference resolution are done (using parses of previous sentences of a multi-sentence caption), statistics are incremented based on the interpretation, and the results are cached. If an interpretation is rejected, search continues. If no satisfactory interpretation can be found, the next-best unexplored word sense is added to the chart and search resumes with it. This is done as many times as necessary. If no interpretation can be found with any word sense, the best pieces of interpretations are assembled for a partial interpretation.

5. Unary statistics

Wordnet is based on traditional printed dictionaries and distinguishes many word senses. A key factor in finding the best sentence interpretation is the relative likelihoods of these word senses. A priori likelihoods can be estimated from the word-sense frequencies ("unary counts") of previously parsed captions. If the word sense has been seen before, its inferred frequency is its observed frequency. These counts stored should include "indirect" counts too, those of all subtypes, so for instance every occurrence of an aircraft counts as an occurrence of a vehicle. Otherwise, if we have not seen a word sense or its subtype before, we infer a frequency of M*K/(1+K) from a neighbor concept if we can, where K is the count of the neighbor and M is a constant reflecting the closeness of the neighbor. M was set from experience to 0.7 for aliases of the word sense; 0.5 for

immediate superconcepts; 0.3 for two-step superconcepts; and 0.8 for links between verbs and verbals. Counts on multiple superconcepts must be summed, as when iron is both an element and a metal. Adjustments are made for complex lexical expressions, like dates and hyphenated words, and for words with very common superconcepts like names of people.

For instance, suppose a sentence contains the word "wing", which has seven Wordnet senses. Sense 1 is an air squadron; sense 4 is a part of an airplane; sense 6 is a part of a building; and sense 7 is a limb of a bird. Sense 4 occurs eleven times in our training and test captions, and sense 6 occurs once. The other senses do not occur. However, sense 1 has an immediate superconcept of "air unit" sense 1, and this does occur twelve times in the captions (albeit never directly, only as subtypes). Sense 7 has a superconcept of "limb" sense 1, which in turn has a superconcept of "extremity" sense 5, which occurs three times in the captions. Hence the likelihood attached to sense 4 is 11, to sense 6 is 1, to sense 1 is 0.5*12/13 = 0.46, and to sense 7 is 0.3*3/4 = 0.22. Senses 2, 3, and 5 get the default likelihood of 0.1.

Initial unary counts were estimated from frequencies of words in the full set of 36,191 NAWC-WD captions, apportioning the count for each word equally among its possible senses (but artificially boosting to 80% of the total the likelihood of senses not nouns, verbs, adjectives or adverbs). After the first caption set was parsed, we eliminated the initial counts and built new counts from the senses in the meaning lists accepted by the trainer. Computation of indirect counts is time-consuming, so recalculation is done only after a caption set is parsed.

6. Binary statistics

Statistical parsing usually exploits the probabilities of strings of successive words in a sentence (Jones and Eisner, 1992; Charniak, 1993). Binary statistics (the counts of the co-occurrence of pairs of word senses) fit naturally with binary parse rules as an estimate of the likelihood of co-occurrence of the two subparses, although more complex theories have been explored (Basili et al, 1996). For example, a parse of "f-18 landing" with the rule "NP -> NP PARTICIPLEPHRASE" should be rated high because F-18s often land, unlike "basement landing" which should only be rated high if parsed as "NG -> ADJ NG". Estimates of co-occurrence probabilities can inherit in a type hierarchy (Rowe, 1985). So if we have insufficient data on how often an F-18 lands, we may have enough on how often an aircraft lands; and if F-18s are typical aircraft, how often F-18s land is estimated by the product of the "aircraft lands" count and the ratio of the count on "F-18" to the count on "aircraft" and its subtypes. Both word senses can be generalized in searching for counts, so we can use statistics on "F-18" and "moving", on "aircraft" and "moving", or on "vehicle" and "moving". We should use the least generalization having sufficient statistics. It makes sense to lump counts for all morphological variants together, so the count on "F-18" and "land" would also cover "the usn f-18s just landed" when interpreted as a noun phrase.

When a binary parse rule combines subparses that are themselves the results of binary parse rules, we use the counts of the syntactically most important words or "headwords". Headwords are the central nouns of noun phrases, the central verbs of verb phrases and clauses, the participles of participial phrases, the prepositions of prepositional phrases, and so on. (Actually, "headwords" can be multiword but atomic concepts, like "World War I".) For instance, "the big f-18 from china lake landing at armitage field" can also be parsed by "NP -> NP PARTICIPLEPHRASE" with the binary count for "f-18" and "landing" used to rate it, since "f-18" is the principal noun and headword of "the big f-18 from china lake", and "landing" is the participle and headword of "landing at armitage field". The restriction of co-occurrence statistical analysis to headwords is consistent with how semantic cases work (a fundamentally binary concept). But this will miss occasional important more-distant relationships in sentences, like between "plane" and "landing" in "the plane that crashed in the Norfolk landing", and it will miss key words that are not syntactic headwords, as "loading" in "reached the loading area". Other researchers are exploring ways to incorporate such knowledge, but our goal was to see how far we could go with a relatively simple approach. Each binary count should be indexed by its corresponding parse rule, so the alternative parse of "f-18 landing" by "NP -> ADJECTIVE GERUND" would have a different count. To help with nominal compounds and appositives, which can be difficult, we further index counts by the relationship postulated between the headwords. For instance, "evaluation facilities" could mean either the facilities are the agent of the evaluation or the evaluation occurs at the facilities.

One advantage of inheritable binary counts is in identification of unknown words. Though we do not exploit this yet, categories for the unknown words can be inferred by their likelihood of accompanying the neighboring word senses. For instance, in "personnel mounting ghw-12 on an f-18", "ghw-12" is likely to be equipment because of the high likelihoods of co-occurrence of equipment terms with "mount" and "on".

Initial binary counts were estimated from frequencies of neighbor-word pairs in the full set of 36,191 NAWC-WD captions, apportioning the count for each word, like the unary counts, equally among its possible senses (but again artificially boosting the likelihood of senses not nouns, verbs, adjectives or adverbs). After the training set of captions was parsed, those statistics were eliminated and new statistics were computed from the accepted sentences. This bootstrapping (Richardson, 1994) was repeated after each caption set was parsed. Counts were incremented for each node of the parse tree for each sentence, as well as for all pairs of superconcepts of the word senses involved. Counts are also computed on the grammar rules used.

The storage for binary counts required careful design because there are many distinct cases and the data is sparse. For instance for just our 616 training and test captions, there were 2,556 distinct word senses (for 3,172 distinct words) and 73,750 binary counts. For the full set of NAWC-WD captions, we estimate we need 23,000 distinct senses (for 29,082 distinct words) and about 3,900,000 binary counts. So we developed a binary counts data structure that uses four search trees indexed on the first word, the part of speech plus word sense of the first word, the second word, and the part of speech plus

word sense of the second word. Various compression techniques could further reduce storage, like omitting counts within a standard deviation of the predicted value (Rowe, 1985). The standard deviation when n is the size of a random subpopulation, N is the size of the population, and A the count for the population, is $\sqrt{A(N-A)(N-n)/nN^2(N-1)}$ (Cochran, 1977).

7. Control of parsing

We use a parsing method similar in approach to that of (Magerman, 1995) and Model 1 of (Collins, 1997), but using different independence assumptions, and computing the more-useful probability of a full semantic interpretation rather than the probability of a parse tree. Our parser uses four factors to rank possible phrase interpretations: (1) the unary counts on the word senses used, (2) the counts on the grammar rules used, (3) the binary counts on the headword senses conjoined in the parse tree, and (4) miscellaneous factors like the inverse of the distance between the two headwords and compatibility of any conjuncts in length and type. It is usual to treat these as if they were independent probabilities (ignoring normalization issues) and multiply them to get the likelihood (weight) for the whole sentence (Fujisaki et al, 1991). For an N-word phrase we can use:

$$\text{weight} = \prod_{i=1}^{N} n(w_i) \prod_{j=1}^{N-1} n(g_j) \prod_{j=1}^{N-1} a(g_j) \prod_{j=1}^{N-1} m(g_j)$$

where $n(w_i)$ is the count of the word sense chosen for the ith word in the phrase, $n(g_j)$ the count of the grammar rule used at the jth (in preorder traversal of the parse tree) binary-rule application in the parse of the phrase, $a(g_j)$ is the degree of association of the headwords of the subphrases joined by the jth binary rule, and $m(g_j)$ are miscellaneous weighting factors.

If we take the negative of the logarithm of this formula, the problem becomes one of finding a minimum-cost sentence interpretation where cost is a sum of factors. Then the challenge of parsing is to estimate which phrase interpretations are most likely to lead to good sentence interpretations, which means estimating cost factors for subphrases not yet known. Since the A* search algorithm is provably optimal and addresses this sort of problem, we use a variant on it. A* would require for each phrase a lower bound on the costs of the remaining factors for the remaining words in the sentence. This is equivalent to finding upper bounds on the factors in the above formula for each remaining word. For the unary-counts factor, we can take the count of the most common sense of a word. For the grammar-rule factor, we can take the count of the most common grammar rule. For the miscellaneous factor, we can take 1 since the factor is computed to have that maximum. But the binary-counts factor is unbounded since the degree of association can vary enormously depending on the corpus. We thus use assume a constant binary-counts cost per remaining word of the sentence, which is equivalent to adding a negative cost constant for each of the words included in the phrase. This gives a revised formula for the likelihood of a phrase:

$$\text{weight} = c^{N-1} \prod_{i=1}^{N} f(w_i) \prod_{j=1}^{N-1} f(g_j) \prod_{j=1}^{N-1} a(g_j) \prod_{j=1}^{N-1} m(g_j)$$

where N is the number of words in the phrase; c is a constant controlling our bias towards longer phrases; $f(w_i)$ is the count of the ith word sense in the sentence divided by the count of the most common sense of that word; $f(g_j)$ is the count of the grammar rule used at the jth binary-rule application divided by the count of the most common grammar rule; $a(g_j)$ is the degree of association of the two headwords of the subphrases joined by the jth binary rule; and $m(g_j)$ is the sum of miscellaneous factors on the jth binary rule. The c is periodically increased when the system has trouble finding a sentence interpretation, and is also adjusted after every sentence interpretation is found to try to keep overall sentence weights between 0.1 and 10.0; these modifications help prevent the parser from getting stuck on sentences with unfamiliar word-sense combinations.

The degree of association between headwords is the ratio of the observed count of the two headwords in this syntactic relationship to expected count. The expected count can be estimated from a log-linear model, the count of this syntactic relationship in the corpus times the proportional frequency of the two word senses in the corpus:
$a(s_j) = (b(w_{j1}, w_{j2})n(t(w_{j1}))n(t(w_{j2})))$ $(\text{b}(\text{t}(\text{w}_{\text{j1}}), t(w_{j2}))n(t(w_{j1}))n(t(w_{j2})))$
where b is the binary frequency, n the unary frequency, j1 the first word sense, j2 the second word sense, and t the topmost generalization of the word sense (which for Wordnet is "entity" sense 1 for nouns and "act" sense 2 for verbs). This is greater than 1 for positively associated words, and less than 1 for negatively associated words. A default of 0.01 is used for word combinations with no statistics.

Take the example sentence "pod on f-4"; Fig. 1 shows the full chart of phrase-interpretation records created. The first argument to each is the index number (and creation order) of the record; the second and third arguments are the starting and ending positions of the phrase in the full sentence; the fourth argument is the syntactic term for the phrase; the fifth the meaning list found; the sixth the backpointers to the component records; and the seventh the weight. Records 1-29 in Fig. 1 cover single words and represent the initial records for the search. For "pod on f-4" we found three Wordnet senses of "pod" as a noun (fruit, animal group, and container) and one domain-dependent verb sense (meaning to put something into a container). Only the third noun sense occurred in the previously seen captions, so it was given a much higher weight; "pod" as a verb got a very small weight since the true verbs are rare in these captions. (The weights of initial records are normalized to enable the constant c to start at 1.) Similarly, "on" can be a location preposition, an orientation preposition, a time preposition, or an adverb, but only the first sense occurred in the previously seen captions. "F-4" is unambiguous and got a weight of 1. Each word sense was then generalized by all possible unary parse rules; so a noun can be an "ng", an "ng" can be an "adj2", and an "adj2" can be an "np". Syntactic categories with multiple generalizations (like "ng"

which can be either an "np" or an "adj2"), split the weight between the generalizations, so "pod-3" as an "np" gets a weight of 0.5.

Then tree-building began. The location interpretation of "on" was chosen first to combine with other agenda items since its weight was highest. Combining it with the word to its right, we got entry 30 for the subphrase "on f-4" with meaning list [on(v6,v12), a_kind_of(v12,'F-4'-0)]. This means some unknown thing v6 is on a v12 which is an instance of an F-4. The computed weight was 0.999 * 0.5 * 1 * 1.018 * 1 = 0.5089 from respectively the weight for this interpretation of "on", the weight for "F-4", a rule-strength weight of 1 (since this is the only rule for these two syntactic categories), a degree of association of 1.018 for "on" and "F-4" (inherited from the degree for "on" and "fighter" sense 4, the aircraft sense), and 1 for the absence of miscellaneous factors. Eventually the search chose entry 30 to work on, and combined it with the three noun senses of "pod" to generate entries 31, 33, and 35, which were immediately generalized to the syntactic category "caption" by a unary parse rule to get entries 32, 34, and 36. Entry 35 got a much higher weight than 31-34 did because 0.5 * 0.509 * 1 * 1.377 * 1 = 0.351 where 0.5 is the weight for "pod" sense 3, 0.509 the weight for "on F-4", 1 the rule weight, 1.377 the degree of association of "pod" sense 3 and "on" (an association in the captions seen), and with no miscellaneous factors. Entry 36 is the final answer, but search continued for a while until all potentially better candidates had been explored.

Fig. 2 shows a longer example comparing MARIE-2 parser output with MARIE-1 output. MARIE-1's output is less precise (without word senses and with very general predicate names), more complex, and shows the effects of overly specialized rules as in the handling of the coordinates. MARIE-1 could not connect sentences, and also erred in identifying a DVT-7 as equipment and "run 2" as the direct object of a test. In general, MARIE-2's meaning lists were significantly more accurate than those of MARIE-1 because it could backtrack to get the best interpretation of a caption, not just an adequate one; MARIE-1 found only one interpretation. The significantly more complex grammatical and semantic features of MARIE-2 also helped.

p(1,1,1,verb,[a_kind_of(v1,pod-100),quantification(v1,plural)],[],0.000544).
p(2,1,1,vg,[a_kind_of(v1,pod-100),quantification(v1,plural)],[1],0.000544).
p(3,1,1,vp2,[a_kind_of(v1,pod-100),quantification(v1,plural)],[2],0.000544).
p(4,1,1,vp,[a_kind_of(v1,pod-100),quantification(v1,plural)],[3],0.000544).
p(5,1,1,noun,[a_kind_of(v2,pod-1)],[],0.015151).
p(6,1,1,ng,[a_kind_of(v2,pod-1)],[5],0.015151).
p(7,1,1,adj2,[a_kind_of(v2,pod-1)],[6],0.007575).
p(8,1,1,np,[a_kind_of(v2,pod-1)],[6],0.007575).
p(9,1,1,noun,[a_kind_of(v3,pod-2)],[],0.015151).
p(10,1,1,ng,[a_kind_of(v3,pod-2)],[9],0.015151).
p(11,1,1,adj2,[a_kind_of(v3,pod-2)],[10],0.007575).
p(12,1,1,np,[a_kind_of(v3,pod-2)],[10],0.007575).
p(13,1,1,noun,[a_kind_of(v4,pod-3)],[],0.999969).
p(14,1,1,ng,[a_kind_of(v4,pod-3)],[13],0.999969).
p(15,1,1,adj2,[a_kind_of(v4,pod-3)],[14],0.499984).
p(16,1,1,np,[a_kind_of(v4,pod-3)],[14],0.499984).
p(17,2,2,locprep,[property(v6,on)],[],0.999995).
p(18,2,2,prep,[property(v6,on)],[17],0.999995).
p(19,2,2,miscprep,[property(v7,orientation)],[],0.002439).
p(20,2,2,prep,[property(v7,orientation)],[19],0.002439).
p(21,2,2,adv,[property(v8,on-150)],[],0.002439).
p(22,2,2,timeprep,[property(v9,during)],[],0.002439).
p(23,2,2,prep,[property(v9,during)],[22],0.002439).
p(24,2,2,miscprep,[property(v10,object)],[],0.019512).
p(25,2,2,prep,[property(v10,object)],[24],0.019512).
p(26,3,3,noun,[a_kind_of(v12,'F-4'-0)],[],0.999833).
p(27,3,3,ng,[a_kind_of(v12,'F-4'-0)],[26],0.999833).
p(28,3,3,adj2,[a_kind_of(v12,'F-4'-0)],[27],0.499916).
p(29,3,3,np,[a_kind_of(v12,'F-4'-0)],[27],0.499916).
p(30,2,3,pp,[on(v6,v12),a_kind_of(v12,'F-4'-0)],[[17,29]],0.508953).
p(31,1,3,np,[a_kind_of(v2,pod-1),on(v2,v12),a_kind_of(v12,'F-4'-0)],[[8,30]],0.002336).
p(32,1,3,caption,[a_kind_of(v2,pod-1),on(v2,v12),a_kind_of(v12,'F-4'-0)],[31], 0.00234).
p(33,1,3,np,[a_kind_of(v3,pod-2),on(v3,v12),a_kind_of(v12,'F-4'-0)], [[12,30]],0.00234).
p(34,1,3,caption,[a_kind_of(v3,pod-2),on(v3,v12),a_kind_of(v12,'F-4'-0)], [33],0.00234).
p(35,1,3,np,[a_kind_of(v4,pod-3),on(v4,v12),a_kind_of(v12,'F-4'-0)], [[16,30]],0.35050).
p(36,1,3,caption,[a_kind_of(v4,pod-3),on(v4,v12),a_kind_of(v12,'F-4'-0)], [35],0.35050).
p(37,1,3,vp2,[a_kind_of(v1,pod-100), quantification(v1,plural),on(v1,v12)
,a_kind_of(v12,'F-4'-0)], [[3,30]],0.00000).
p(38,1,3,vp,[a_kind_of(v1,pod-100),quantification(v1,plural),on(v1,v12),
a_kind_of(v12,'F-4'-0)],[37],0.000000).
p(40,2,3,pp,[object(v10,v12),a_kind_of(v12,'F-4'-0)],[[24,29]],0.009611).

Figure 1: Chart resulting from parse of "pod on f-4".

Input 215669:

"tp 1314. a-7b/e dvt-7 (250 keas) escape system (run 2). synchro firing at 1090' n x 38' w. dummy just leaving sled."

Meaning list computed by MARIE-2:

[a_kind_of(v1,"TP-1314"-0), during(v3,v1), a_kind_of(v3,"escape system"-0),
during(v3,v4),
a_kind_of(v4,"RUN 2"-0), agent(v8,v3), a_kind_of(v8,"DVT-7"-0),
measurement(v8,v2),
a_kind_of(v2,"250 keas"-0), quantity(v2,250), units(v2,keas), object(v8,v5),
a_kind_of(v5,"A-7B/E"-0),
during(v141,v1), a_kind_of(v141,launch-2), property(v141,synchronous-51),
at(v141,v95),
a_kind_of(v95,place-8), part_of(v45,v95), part_of(v46,v95), a_kind_of(v45,"1090" n"-0),
quantity(v45,1090), units(v45,"latitude-minute"-0), a_kind_of(v46,"38" w"-0),
quantity(v46,38),
units(v46,"longitude-minute"-0), during(v999,v1), a_kind_of(v999,dummy-3),
agent(v1012,v999),
a_kind_of(v1012,leave-105), tense(v1012,prespart), property(v1012,just-154),
object(v1012,v1039),
 a_kind_of(v1039,sled-1)]

Superconcept information for the word senses:

"TP-1314"-0 -> test-3, "escape system"-0 -> system-8, "DVT-7"-0 -> test-3,
"250 keas"-0 -> number-7, "A-7B/E"-0 -> fighter-4, "RUN 2"-0 -> run-1, shoot-109 -> discharge-105,
place-8 -> "geographic area"-1, "1090' n"-0 -> number-6, "38' w"-0 -> number-6,
synchronous-51 -> "at the same time"-51, dummy-3 -> figure-9, leave-105 -> go-111,
sled-1 -> vehicle-1

Meaning list computed by MARIE-1:

[inst('noun(215669-4-2)',sled), inst('noun(215669-2-a68)','A-7B/E'),
attribute('noun(215669-2-a67)',part_of('noun(215669-2-a68)')),
inst('noun(215669-2-a67)','DVT-7'), agent('prespart(215669-4-1)',obj('noun(215669-4-1)')),
source('prespart(215669-4-1)',obj('noun(215669-4-2)')),
activity('prespart(215669-4-1)',depart), attribute('coordinate(215669-3-1)','1090 " N x 38 " W'),
inst('coordinate(215669-3-1)',coordinate), attribute('noun(215669-2-6)','2'),
theme('noun(215669-2-6)',obj('noun(215669-2-5)')), inst('noun(215669-2-6)',run),

theme('noun(215669-2-5)',obj('noun(215669-2-a67)')), quantity('noun(215669-2-5)',keas('250')),
inst('noun(215669-2-5)',test), attribute('noun(215669-3-1)',synchronous),
location('noun(215669-3-1)',at('coordinate(215669-3-1)')),
activity('noun(215669-3-1)',launch), inst('noun(215669-4-1)',dummy),
attribute('noun(215669-1-1)','1314'), inst('noun(215669-1-1)','test plan')]

Figure 2: Example parser output, plus superconcepts for the word senses used.

8. Experiments

We used 616 captions in four caption sets comprising 1009 sentences in our experiments (see Table 3). The first set was the 217 captions handled by MARIE-1 (Guglielmo and Rowe, 1996). These were from the NAWC-WD Photo Lab, and were created by taking a random sample of supercaptions (captions for sets of photographs) and transcribing the captions written on photograph folders for all their component photographs. Since captions for a supercaption are closely related, there was significant redundancy. Some correction of syntax errors was done as explained in the earlier paper. Statistics for parsing this caption set were estimated as described in sections 5 and 6. After training on this set, we calculated statistics and ran caption set 2 using them. Set 2 captions were a different random sample of 108 supercaptions with little redundancy. No manual correction of the captions was done, and they were difficult to parse. Caption sets 3 and 4 were 172 and 119 manually-extracted captions constituting nearly all the captioned images available on the NAWC-WD World Wide Web site in August 1998. These were less technical and used some new grammatical constructs like conjunctive adjectives and relative clauses, but were mostly not difficult to parse. Sets 3 and 4 were not used for initial lexicon construction and so required additional lexicon entries. After each caption set we updated our statistics using the new results, so set 1 provided statistics for testing set 2, sets 1 and 2 for testing set 3, and sets 1, 2, and 3 for testing set 4. We finally had counts for 1931 distinct word senses and 4018 word-sense pairs. Development and testing took about a man-year of work including development of the parser.

Table 3 : Overall statistics on the four caption sets (numbers in parentheses are the occurrence rate per caption word).

	Caption set 1 (training)	Caption set 2 (test/training)	Caption set 3 (test/training)	Caption set 4 (test)
Number of new captions	217	108	172	119
Number of new sentences	444	219	218	128
Total number of words in new captions	4488	1774	1535	1085
Number of distinct words in new captions	939 (.2092)	900 (.5073)	677 (.4410)	656 (.6046)
Number of new lexicon entries required	c. 150 (.0334)	106 (.0598)	139 (.0906)	53 (.0488)
Number of new word senses used	929 (.2070)	728 (.4104)	480 (.3127)	416 (.3834)
Number of new sense pairs used	1860 (.4144)	1527 (.8608)	1072 (.6983)	795 (.7327)
Number of lexical-processing changes required	c. 30 (.0067)	11 (.0062)	8 (.0052)	7 (.0065)
Number of syntactic-rule changes or additions	35 (.0078)	41 (.0231)	29 (.0189)	10 (.0092)
Number of case-definition changes or additions	57 (.0127)	30 (.0169)	16 (.0104)	3 (.0028)
Number of semantic-rule changes or additions	72 (.0161)	57 (.0321)	26 (.0169)	14 (.0129)

On each caption sentence, we forced the system to backtrack and try again until it found the best possible interpretation according to the trainer's (the author's) judgment. To guide it, with each sentence interpretation generated, the system permitted the trainer to rule out one particular error (only one per try, to get useful statistics on the number of tries). So the trainer could tell the parser to rule out "sidewinder" sense 2, part-whole relationships, past participles, conjunctive connections, or even a specific predicate expression saying an aircraft sense 1 possessed a sidewinder sense 2. When bugs in the parser were occasionally found, processing was aborted, the bugs were fixed, and the sentence was run again.

Fig. 3 shows the distribution of total parse times (including all tries for each sentence) in CPU seconds (using semi-compiled Quintus Prolog) for the 1009 caption sentences as a function of sentence length in number of words; the axes display the natural logarithms. (Bear in mind this implementation was only semi-compiled, and could be speeded up significantly.) Significant sentence variation is apparent. Fig. 4 shows the logarithm of the geometric mean of parse CPU time as a function of the logarithm of sentence length; the textured line is the first caption set, the dashed line is the second, and the solid line is

the third and fourth sets. (We combine statistics for the third and fourth sets since the statistics were very similar.) It can be seen that parse time increased on the second caption set due to the increased parser complexity after debugging, but not subsequently. This suggests that initialization effects are now fading and performance will remain constant or improve with further training captions. Also note all the curves have a linear trend in these log-log plots; least-squares regression gave us a rough fit of $0.0858n^{2.876}$ for n is the number of words in the sentence, with degree of significance 0.531.

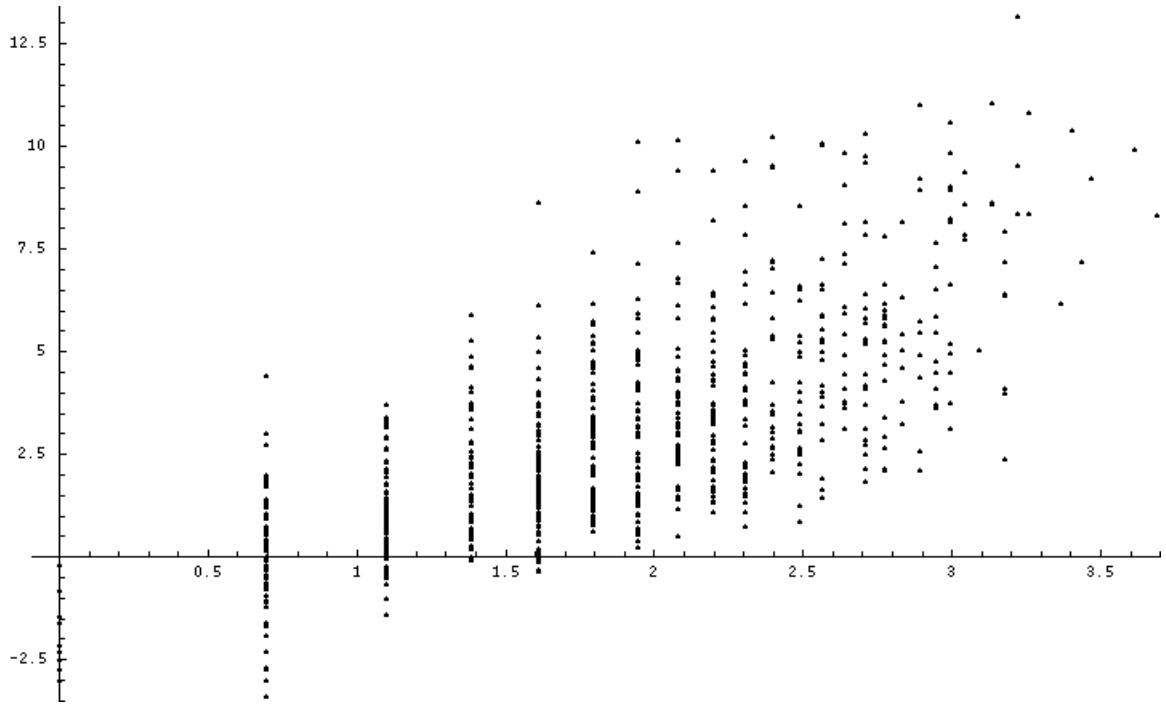


Figure 3: Logarithm of parse CPU time versus logarithm of sentence length for all captions.

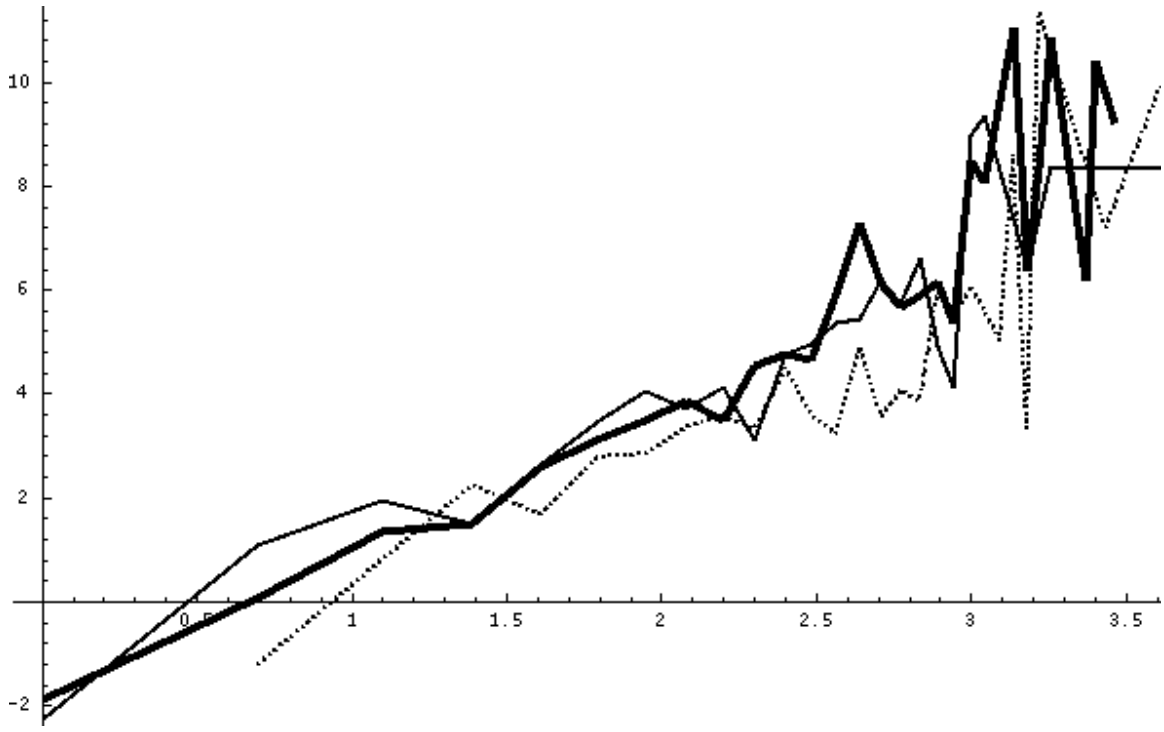


Figure 4: Natural logarithm of average parse CPU time in seconds versus sentence length (hatched line is for set 1, dotted line is for set 2, and solid line is for sets 3 and 4).

Successful parses were found for all but two ungrammatical sentences (and two very long sentences had to be split to parse in a reasonable amount of time). Thus to measure parse accuracy we use the number of tries before the best sentence interpretation was found. As mentioned above, only one word sense or relationship can be corrected per try, so the number of tries is a rough metric of the number of errors in the first-attempt meaning list. Fig. 5 plots the logarithm of the number of tries as a function of the logarithm of sentence length. Fitting this data, we got the formula $1.809n^{1.668}$, where n is the number of words in the sentence, with degree of significance 0.550. The shape of the curve is similar to that for CPU time, but sets 3 and 4 show a little improvement over set 2, suggesting further improvements to come.

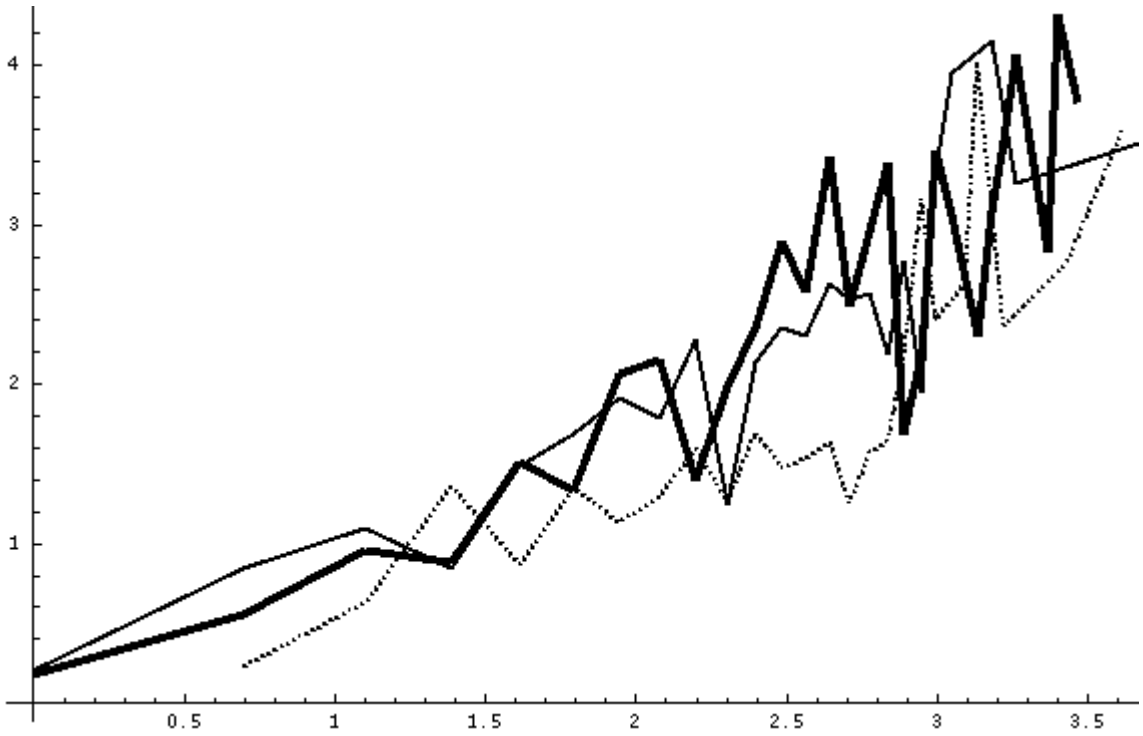


Figure 5: Natural logarithm of average number of tries versus logarithm of sentence length (hatched line is for set 1, dotted line is for set 2, and solid line is for sets 3 and 4).

To study the influence of key factors on the parser, we did more detailed experiments with eight representative captions from caption set 3 shown in Table 4. Table 5 shows parse CPU time and number of required tries before obtaining the best interpretation for each caption. The first pair of numbers are during the test run using statistics from the first two caption sets. The second pair of numbers are from rerunning after including statistics from all of set 3. The new statistics clearly help but not much on shorter sentences; the best improvement was for the last sentence with hints that the unusual constructs of gerunds and prepositional adverbs were being used. The third pair of numbers in Table 5 are for conditions like the second except without any binary word-sense statistics. Inheritance of binary statistics slows some sentences a little, but help some longer sentences like the seventh with its many nominal compounds. The fourth pair of numbers are for conditions like the second except without unary statistics, so that all word senses were rated equally likely. This hurts performance significantly, suggesting that learning the common word senses is a key to learning this dialect.

Table 4: Example sentences.

NO.	CAPTION
1	pacific ranges and facilities department, sled tracks.
2	airms, pointer and stabilization subsystem characteristics.
3	vacuum chamber in operation in laser damage facility.
4	early fleet training aid: sidewinder 1 guidance section cutaway.
5	awaiting restoration: explorer satellite model at artifact storage facility.
6	fae i (cbu-72), one of china lake's family of fuel-air explosive weapons.
7	wide-band radar signature testing of a submarine communications mast in the bistatic anechoic chamber.
8	the illuminating antenna is located low on the vertical tower structure and the receiving antenna is located near the top.

Table 5: Parse CPU time and number of tries until the best interpretation is found for the eight example sentences under difference circumstances.

Sentence number	Sentence length	Training time	Training tries	Final time	Final tries	No-binary time	No-binary tries	No-unary time	No-unary tries
1	8	27.07	13	17.93	5	8.27	5	60.63	19
2	7	70.27	10	48.77	9	94.62	14	124.9	23
3	8	163.0	19	113.1	19	202.9	23	2569.0	22
4	9	155.2	9	96.07	3	63.95	8	229.3	22
5	10	86.42	8	41.02	3	49.48	6	130.6	30
6	15	299.3	11	65.78	7	68.08	5	300.4	15
7	15	1624.0	24	116.5	5	646.0	12	979.3	25
8	20	7825.0	28	35.02	2	35.60	3	>50000	-

9. Queries

We also built a natural-language query capability for captions using this parser. An English query obtained from the user is parsed and interpreted. Its variables are given special names different from caption-variable names, and tense and number markings are eliminated. "Coarse-grain matching" is then done to find captions mentioning the concept types in the query. This requires a full index on types mentioned in captions. Caption candidates passing the coarse-grain match are then subject to "fine-grain matching" of the full query meaning list to the full caption meaning list. This is a standard nondeterministic match with backtracking, and must be an exact match. We assume that

the correct interpretation of a query is the highest-ranking interpretation that has a non-null fine-grain match, and we backtrack to automatically generate interpretations until we find one. Thus our parser need not be perfect at finding the best interpretation first; but the less accurate it is, the more time it requires. Thumbnail-sized pictures corresponding to the matched captions are then displayed. High accuracy was shown for this phase of MARIE-1, so we did not test it for MARIE-2.

10. Conclusions

We have attempted a difficult task of parsing substantial sentences in a raw specialized real-world dialect with unusual new word senses, creative syntax, and errors. Our results show it can be done, and statistical parsing helps, but that significant setup work is required. We reached the point after 1009 sentences where the number of tries required to get the best interpretation of a sentence decreased, which is encouraging, but the total parse time remained constant. We suspect this reflects a tradeoff of smarter processing with the added effort in handling increasing numbers of special cases, and could be improved by a more efficient encoding. Nonetheless, parse time may improve with further sentences anyway. And it should likely improve with preliminary sense disambiguation like that of (Leacock, Chodorow, and Miller, 1998). We hope our experiments will provide helpful ideas to other researchers addressing the many specialized technical dialects for which automated understanding can be valuable.

11. References

Basili, R., Pazienza, M., and Velardi, P. A shallow syntactic analyzer to extract word associations from corpora. *Literary and Linguistic Computing, 7*, 2 (1992), 114-124.

Basili, R., Pazienza, M. T., and Velardi, P. An empirical symbolic approach to natural language processing. *Artificial Intelligence, 85* (1996), 59-99.

Brill, E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21, 4 (December 1995), 543-565.

Charniak, E. *Statistical Language Learning*. Cambridge, MA: MIT Press, 1993.

Cochran, W. G. *Sampling Techniques, Third Edition*. New York: Wiley, 1977.

Collins, M. Three generative, lexicalised models for statistical parsing. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, July 7-12, 1997, 16-23.

Fujisaki, T., Jelinek, F., Cocke, J., Black, E., and Nishino, T. A probabilistic parsing method for sentence disambiguation. In *Current issues in parsing technology*, ed. Tomita, M., Boston: Kluwer, 1991.

Grosz, B., Appelt, D., Martin, P. and Pereira, F. TEAM: An experiment in the design of transportable natural language interfaces. *Artificial Intelligence, 32* (1987), 173-243.

Guglielmo, E. and Rowe, N. Natural language retrieval of images based on descriptive captions. *ACM Transactions on Information Systems, 14*, 3 (May 1996), 237-267.

Jones, M. and Eisner, J. A probabilistic parser applied to software testing documents. Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, July 1992, 323-328.

Krovez, R. and Croft, W. B. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems, 10*, 2 (April 1992), 115-141.
Leacock, C., Chodorow, M., and Miller, G. Using corpus statistics and Wordnet relations for sense identification. *Computational Linguistics, 24*, (March 1998), 147-165.
Magerman, D. Statistical decision-tree models for parsing. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA USA, June 26-30, 1995, 276-283.
Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. Five papers on Wordnet. *International Journal of Lexicography, 3*, 4 (Winter 1990).
Rau, L. Knowledge organization and access in a conceptual information system. *Information Processing and Management, 23*, 4 (1988), 269-284.
Richardson, S. D. Bootstrapping statistical processing into a rule-based natural language parser. Proceedings of The Balancing Act: Combining Symbolic and Statistical Approaches to Language, Association for Computational Linguistics, Las Cruces, NM, July 1994, 96-103. Also Microsoft Technical Report MSR-TR-95-48.
Rowe, N. Antisampling for estimation: an overview. *IEEE Transactions on Software Engineering, SE-11*, 10 (October 1985), 1081-1091.
Rowe, N. and Laitinen, K. Semiautomatic disabbreviation of technical text. *Information Processing and Management, 31*, 6 (1995), 851-857.
Sembok, T. and van Rijsbergen, C. SILOL: A simple logical-linguistic document retrieval system. *Information Processing and Management, 26*, 1 (1990), 111-134.
Silvester, J. P., Genuardi, M. T., and Klingbiel, P. H. Machine-aided indexing at NASA. *Information Processing and Management, 30*, 5 (1994), 631-645.

This work was sponsored by DARPA as part of the I3 Project under AO 8939, by the Army Artificial Intelligence Center, and by the U. S. Naval Postgraduate School under funds provided by the Chief for Naval Operations. Thanks to Albert Wong and Eugene Guglielmo for programming help.

Understanding Navy technical language via statistical parsing

Neil C. Rowe
U.S. Naval Postgraduate School
ncrowe@nps.edu

Example sentence for parsing (from NAWC-WD)

- *an/apq-89 xan-1 radar set in nose of t-2 buckeye modified aircraft bu# 7074, for flight evaluation test. 3/4 overall view of aircraft on runway.*
- Note two noun phrases terminated with periods; the first describes the photographic subject and the second the picture itself.
- Note complex nominal compounds, "an/apq-89 xan-1 radar set" and "t-2 buckeye modified aircraft bu# 7074".
- Domain knowledge: "an/apq-89" is a radar, "xan-1" a version number, "t-2" an aircraft, "buckeye" slang for a T-2, "modified" a conventional adjective, and "bu# 7074" as an aircraft code ID.

Example of domain-specific word senses

- *graphics presentation tid progress 76. sea site update, wasp head director and hawk screech/sun visor radars. top portion only, excellent.*
- “Wasp", "hawk", "screech", and "sun visor" should not be interpreted in their common English word senses, but as equipment terms.
- “Progress 76" means "progress in 1976”.
- “Excellent" refers to the quality of the picture.
- The "head director" is not a person but a guidance system.
- The "sea site" is a dry lakebed flooded with water to a few inches.

Example with abbreviations and misspellings

- *aerial low oblique, looking s from inyodern rd at main gate down china lake bl to bowman rd. on l, b to t, water reservoirs, trf crcl, pw cmpnd, vieweg school, capehart b housing, burroughs hs, cimarron gardens, east r/c old duplex stor. lot. on r, b to t, trngl, bar s motel, arrowsmith, comarco, hosp and on to bowman rd.*
- "Trf crcl" is "traffic circle", "trngl" is "triangle", "capehart b" is "capehart base", but "b to t" is "bottom to top".
- "Vieweg" which looks like a misspelling is actually a person name, but "inyodern" should be "inyokern", a nearby town.

Our approach

- Use Wordnet for the basic lexicon.
- Enhance the lexicon with codeword formats and 2000 explicitly written lexicon entries (in the 36,000 captions).
- Use bottom-up chart parser (one and two-replacement rules only) with statistics-based ranking.
- Assign semantics through a set of general-purpose semantic rules with case constraints.
- Rank phrase interpretations as product of a priori probabilities of word senses, probability of the syntax, and probability of the co-occurrence of the two headwords for each parse-tree node.

Rules used in parsing and their counts in the corpus

Rule	Frequency	Example
adj2 + ng = ng	2551	"Navy" + "aircraft"
b_prtp + np = prtp2	122	"testing" + "the seat"
art2 + ng = np	288	"the" + "naval aircraft"
adv + participle = a_prtp	28	"just" + "loaded"
noun + numeric = ng	81	"test" + "0345"
timeprepx + np = pp	82	"during" + "the test"
locprepx + np = pp	710	"on" + "the ground"
miscprepx + np = pp	654	"with" + "instrument pod"
np + pp = np	1241	"Navy aircraft" + "during testing"
np + prtp = np	306	"a crewman" + "loading the pod"
vg + np = vp2	53	"loads" + "the instrument pod"
np + vp = snt	53	"a crewman" + "loads the pod"
vp2 + pp = vp2	25	"loads" + "on the aircraft"
adv + pp = pp	24	"just" + "below the aircraft"
conj + np = cj_np	167	"and" + "aircraft"
np + cj_np = np	195	"sled" + "and dummy"
np + c_aps = np	60	"the aircraft" + ", F-18"
ng + aps = ng	195	"aircraft" + "(F-18)"
np + aps = np	54	"the aircraft" + "(F-18)"
np + c_np = np	55	"the aircraft" + ", the F-18"
np_c + np = np	57	"sled," + "dummy" [comma fault]
prtp2 + pp = prtp2	155	"just loaded" + "on aircraft"
infinmarker + vp = ip	19	"to" + "load"
conj + vp = cj_vp	3	"and" + "loads the aircraft"
snt + cj_snt = snt	3	"crewmen load" + "and officer directs"

Statistics on the corpus

Description	Count
Captions	36,191
Words	610,182
Distinct words	29.082
Subset with entries in Wordnet	6,729
Word senses for the Wordnet words	14,676
Reused from MARIE-1	770
Written explicitly for MARIE-2	1,763
Morphological variants	2,335
Numbers	3,412
Person names	2,791
Place names	387
Manufacturer names	264
Defined-code prefixes	3,256
Defined-code prefixes in these	947
Identifiable special formats	10,179
Misspellings	1,174
Abbreviations	1,093
Assumed equipment names	1,876
Aliases of above Wordnet words	20,299
Alias senses added	16,712
Superconcept aliases	35,976
Total number of word senses handled	69,447

Example entries in parse chart of "pod on f-4".

p(1,1,1,verb, [a_kind_of(v1,pod-100), quantification(v1,plural)],[],0.0005).

p(5,1,1,noun, [a_kind_of(v2,pod-1)],[],0.015151).

p(9,1,1,noun, [a_kind_of(v3,pod-2)],[],0.015151).

p(13,1,1,noun, [a_kind_of(v4,pod-3)],[],0.999969).

p(17,2,2,locprep, [property(v6,on)], [],0.999995).

p(19,2,2,miscprep, [property(v7, orientation)],[],0.002439).

p(20,2,2,prep, [property(v7, orientation)],[19],0.002439).

p(21,2,2,adv, [property(v8,on-150)],[],0.002439).

p(22,2,2,timeprep, [property(v9,during)],[],0.002439).

p(26,3,3,noun,[a_kind_of(v12,'F-4'-0)],[],0.999833).

p(27,3,3,ng,[a_kind_of(v12,'F-4'-0)],[26],0.999833).

p(30,2,3,pp,[on(v6,v12),a_kind_of(v12, 'F-4'-0)],[[17,29]],0.508953).

p(31,1,3,np,[a_kind_of(v2,pod-1),on(v2,v12),a_kind_of(v12,'F-4'-0)],[[8,30]],0.002336).

p(32,1,3,caption,[a_kind_of(v2,pod-1),on(v2,v12),a_kind_of(v12,'F-4'-0)],[31], 0.00234).

p(33,1,3,np,[a_kind_of(v3,pod-2),on(v3,v12),a_kind_of(v12,'F-4'-0)], [[12,30]],0.00234).

p(34,1,3,caption,[a_kind_of(v3,pod-2),on(v3,v12),a_kind_of(v12,'F-4'-0)], [33],0.00234).

Example computed meaning list

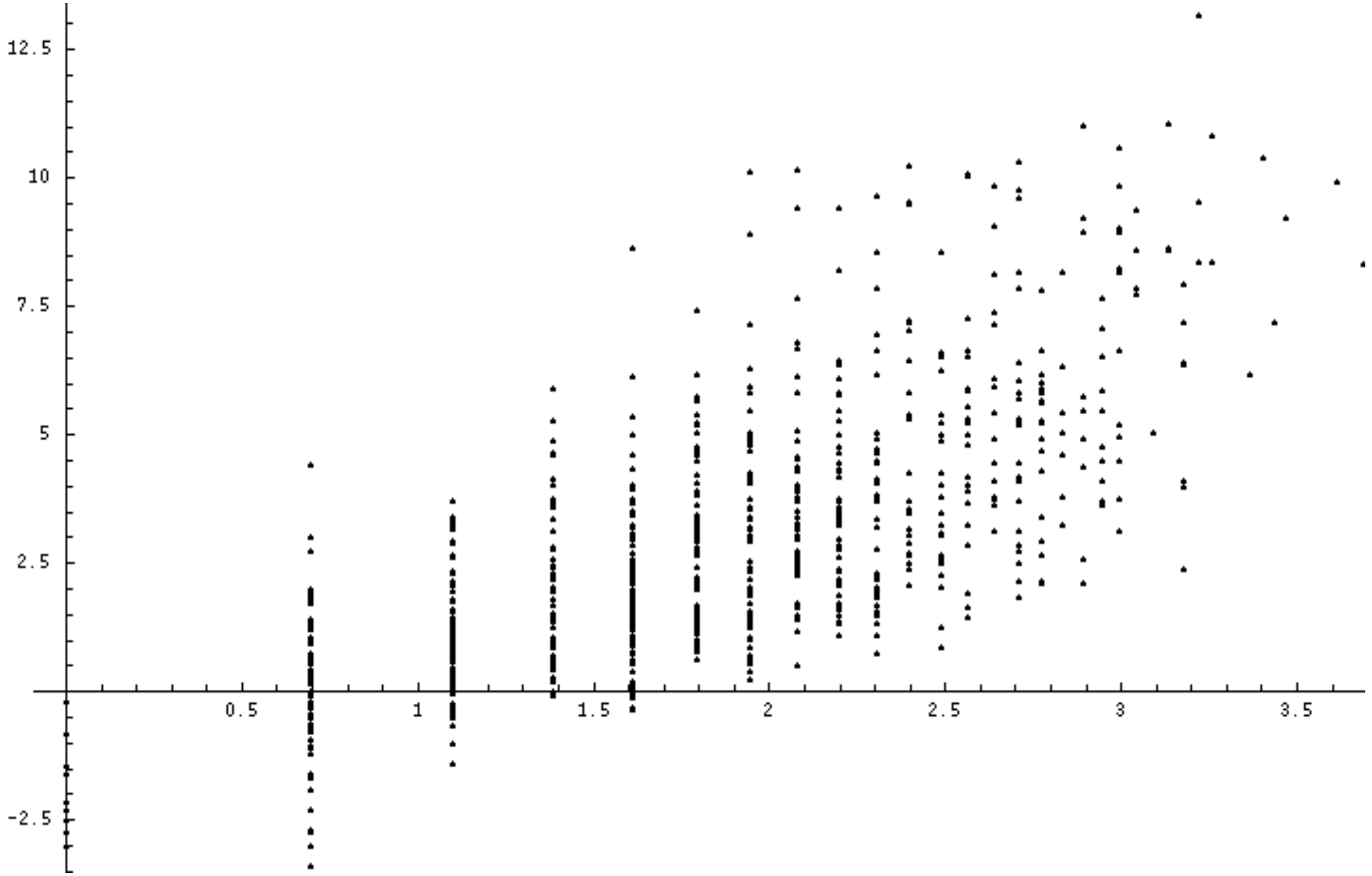
"tp 1314. a-7b/e dvt-7 (250 keas) escape system (run 2). synchro firing at 1090' n x 38' w. dummy just leaving sled."

[a_kind_of(v1,"TP-1314"-0), during(v3,v1), a_kind_of(v3,"escape system"-0), during(v3,v4), a_kind_of(v4,"RUN 2"-0), agent(v8,v3), a_kind_of(v8,"DVT-7"-0), measurement(v8,v2), a_kind_of(v2, "250 keas"-0), quantity(v2,250), units(v2,keas), object(v8,v5), a_kind_of(v5,"A-7B/E"-0), during(v141,v1), a_kind_of(v141,launch-2), property(v141,synchronous-51), at(v141,v95), a_kind_of(v95,place-8), part_of(v45,v95), part_of(v46,v95), a_kind_of(v45,"1090" n"-0), quantity(v45,1090), units(v45,"latitude-minute"-0), a_kind_of(v46,"38" w"-0), quantity(v46,38), units(v46,"longitude-minute"-0), during(v999,v1), a_kind_of(v999,dummy-3), agent(v1012,v999), a_kind_of(v1012,leave-105), tense(v1012,prespart), property(v1012,just-154), object(v1012,v1039), a_kind_of(v1039,sled-1)]

Statistics on the training and test runs

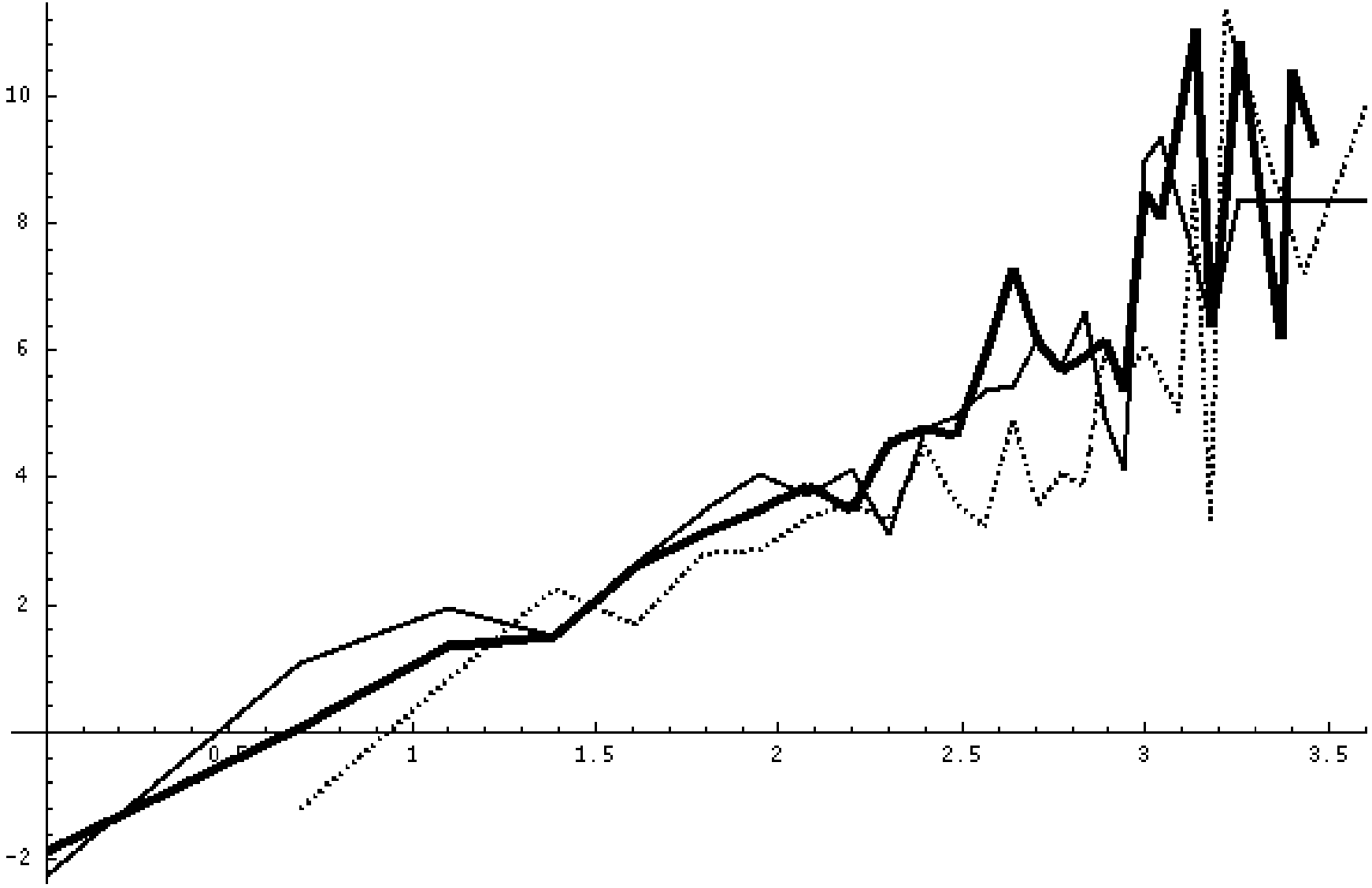
	Caption set 1 (training)	Caption set 2 (test/training)	Caption set 3 (test/training)	Caption set 4 (test)
Number of new captions	217	108	172	119
Number of new sentences	444	219	218	128
Total number of words in new captions	4488	1774	1535	1085
Number of distinct words in new captions	939 (.2092)	900 (.5073)	677 (.4410)	656 (.6046)
Number of new lexicon entries required	c. 150 (.0334)	106 (.0598)	139 (.0906)	53 (.0488)
Number of new word senses used	929 (.2070)	728 (.4104)	480 (.3127)	416 (.3834)
Number of new sense pairs used	1860 (.4144)	1527 (.8608)	1072 (.6983)	795 (.7327)
Number of lexical-processing changes required	c. 30 (.0067)	11 (.0062)	8 (.0052)	7 (.0065)
Number of syntactic-rule changes or additions	35 (.0078)	41 (.0231)	29 (.0189)	10 (.0092)
Number of case-definition changes or additions	57 (.0127)	30 (.0169)	16 (.0104)	3 (.0028)
Number of semantic-rule changes or additions	72 (.0161)	57 (.0321)	26 (.0169)	14 (.0129)

Log of processing time versus sentence length



Trend of processing time versus sentence length

(Hatched = set 1, dotted = set 2, solid = sets 3 & 4)



8 sentences used for comparative tests

NO.	CAPTION
1	pacific ranges and facilities department, sled tracks.
2	airms, pointer and stabilization subsystem characteristics.
3	vacuum chamber in operation in laser damage facility.
4	early fleet training aid: sidewinder 1 guidance section cutaway.
5	awaiting restoration: explorer satellite model at artifact storage facility.
6	fae i (cbu-72), one of china lake's family of fuel-air explosive weapons.
7	wide-band radar signature testing of a submarine communications mast in the bistatic anechoic chamber.
8	the illuminating antenna is located low on the vertical tower structure and the receiving antenna is located near the top.

Comparative results for 8 test sentence

This shows that both unary and binary word-occurrence information helps, as does training.

Sentence number	Sentence length	Training time	Training tries	Final time	Final tries	No-binary time	No-binary tries	No-unary time	No-unary tries
1	8	27.07	13	17.93	5	8.27	5	60.63	19
2	7	70.27	10	48.77	9	94.62	14	124.9	23
3	8	163.0	19	113.1	19	202.9	23	2569.0	22
4	9	155.2	9	96.07	3	63.95	8	229.3	22
5	10	86.42	8	41.02	3	49.48	6	130.6	30
6	15	299.3	11	65.78	7	68.08	5	300.4	15
7	15	1624.0	24	116.5	5	646.0	12	979.3	25
8	20	7825.0	28	35.02	2	35.60	3	>5000 0	-